Project-Team SHAMAN

A Symbolic and Human-Centric View of Data Management

Lannion

Activity Report

2014

1 Team

Head of the team

Olivier Pivert, Professor, Enssat

Administrative assistant

Joëlle Thépault, Enssat, 20%
Nelly Vaucelle, Enssat, 10% (up to August 2014)
Vincent Chevrette, Enssat, 10% (since September 2014)

Université Rennes 1 personnel

François Goasdoué, Professor, Enssat
Hélène Jaudoin, Associate Professor, Enssat
Ludovic Liétard, Associate Professor, HdR, IUT Lannion
Daniel Rocacher, Professor, Enssat
Grégory Smits, Associate Professor, IUT Lannion
Virginie Thion, Associate Professor, Enssat

PhD students

Katia Abbaci, Région Bretagne grant and ANR grant, (Nov 2009 – Aug. 2014)
Sara El Hassad, Région Bretagne grant and Lannion Trégor Communauté grant, since October 2014
William Correa Beltran, Région Bretagne grant and Conseil Général 22 grant, since October 2012
Aurélien Moreau, DGA contract, since November 2014
Olfa Slama, DGA contract, since November 2014

Master students

Aurélien Moreau, Enssat, March-July 2014

2 Overall Objectives

In database research, the last two decades have witnessed a growing interest in preference queries on the one hand, and uncertain databases on the other hand.

Motivations for introducing preferences inside database queries are manifold. First, it has appeared to be desirable to offer more expressive query languages that can be more faithful to what a user intends to say. Second, the introduction of preferences in queries provides a basis for rank-ordering the retrieved items, which is especially valuable in case of large sets of items satisfying a query. Third, on the contrary, a classical query may also have an empty set of answers, while a relaxed (and thus less restrictive) version of the query might be matched by items in the database.

Approaches to database preference queries may be classified into two categories according to their qualitative or quantitative nature. In the qualitative approach, preferences are defined through binary preference relations. Among the representatives of this family of approaches, let us mention an approach based on CP-nets, and those relying on a dominance relation, e.g. Pareto order, in particular Skyline queries. In the quantitative approach, preferences are expressed quantitatively by a monotone scoring function (the overall score is positively correlated with partial scores). Since the scoring function associates each tuple with a numerical score, tuple t_1 is preferred to tuple t_2 if the score of t_1 is higher than the score of t_2. Well-known representatives of this family of approaches are top-k queries, and *fuzzy-set-based approaches*. The team Shaman particularly studies the latter, and the line followed is to focus on:

1. various types of flexible conditions, including non-trivial ones,
2. the semantics of such conditions from a user standpoint,
3. the design of query languages providing flexible capabilities in a relational setting.

Basically, a fuzzy query involves linguistic terms corresponding to gradual predicates, i.e., predicates which are more or less satisfied by a given (attribute) value. In addition, these various terms may have different degrees of importance, which means that they may be connected by operators beyond conjunction and disjunction. For instance, in the context of a search for used vehicles, a user might say that he/she wants a *compact* car *preferably French*, with a *medium* mileage, *around* 6 k$, whose color is *as close as possible* to light grey or blue. The terms appearing in this example must be specified, which requires a certain theoretical framework. For instance, one may think that "*preferably* French" corresponds to a complete satisfaction for French cars, a lower one for Italian and Spanish ones, a still smaller satisfaction for German cars and a total rejection for others. Similarly, "*medium* mileage" can be used to state that cars with less than 40000 km are totally acceptable while the satisfaction decreases as the mileage goes up to 75000 km which is an upper bound. Moreover, it is likely that some of the conditions are more important than others (e.g., the price with respect to the color). In such a context, answers are ordered according to their overall compliance with the query, which makes a major difference with respect to usual queries.

In the previous example, conditions are fairly simple, but it turns out that more complex ones can also be handled. A particular attention is paid to conditions calling on aggregate functions together with gradual predicates. For instance, one may look for departments where *most* employees are *close* to retirement, or where the average salary of *young* employees is *around* $2500. Such statements have their counterpart in regular query language, such as SQL, and the specification of their semantics, when gradual conditions come into play, is studied in the project.

Along this line, the ultimate goal of the project is to introduce gradual predicates inside database query languages, thus providing flexible querying capabilities. Algebraic languages as well as more user-oriented languages are under consideration in both the original and extended relational settings.

As to the second topic mentioned at the beginning of this introduction, i.e., uncertain databases, it already has a rather long history. Indeed, since the late 70s, many authors have

made diverse proposals to model and handle databases involving uncertain or incomplete data. In particular, the last two decades have witnessed a profusion of research works on this topic. The notion of an uncertain database covers two aspects: i) attribute uncertainty: when some attribute values are ill-known; ii) existential uncertainty: when the existence of some tuples is itself uncertain. Even though most works about uncertain databases consider probability theory as the underlying uncertainty model, some approaches rather rely on possibility theory. The issue is not to demonstrate that the possibility-theory-based framework is "better" than the probabilistic one at modeling uncertain databases, but that it constitutes an interesting alternative inasmuch as it captures a different kind of uncertainty (of a subjective, nonfrequential, nature). A typical example is that of a person who witnesses a car accident and who does not remember for sure the model of the car involved. In such a case, it seems reasonable to model the uncertain value by means of a possibility distribution, e.g., {1/Mazda, 1/Toyota, 0.7/Honda} rather than with a probability distribution which would be artificially normalized. In contrast with probability theory, one expects the following advantages when using possibility theory:

- the qualitative nature of the model makes easier the elicitation of the degrees attached to the various candidate values;
- in probability theory, the fact that the sum of the degrees from a distribution must equal 1 makes it difficult to deal with incompletely known distributions;
- there does not exist any probabilistic logic which is complete and works locally as possibilistic logic does: this can be problematic in the case where the degrees attached to certain pieces of data must be automatically deduced from those attached to some other pieces of data (e.g., when data coming from different sources are merged into a single database).

A recent research topic in Shaman concerns flexible data integration systems. One considers a distributed database environment where several data sources are available. An extreme case is that of a totally decentralized P2P system. An intermediary situation corresponds to the case where several global schemas are available and where the sources can be accessed through views defined on one of these schemas (LAV approach). The problem consists in handling a user query (possibly involving preferences conveyed by fuzzy terms) so as to forward it (or part of it) to the relevant data sources, after rewriting it using the views. The overall objective is thus to define flexible query rewriting techniques which take into account both the approximate nature of the mappings and the graded nature of the initial query. A large scale environment is aimed, and the performance aspect is therefore crucial in such a context.

3 Scientific Foundations

The project investigates the issues of flexible queries against regular databases as well as regular queries addressed to databases involving imprecise data. These two aspects make use of two close theoretic settings: fuzzy sets for the support of flexibility and possibility theory for the representation and treatment of imprecise information.

3.1 Fuzzy sets

Fuzzy sets were introduced by L.A. Zadeh in 1965 [Zad65] in order to model sets or classes whose boundaries are not sharp. This is particularly the case for many adjectives of the natural language which can be hardly defined in terms of usual sets (e.g., high, young, small, etc.), but are a matter of degree. A fuzzy (sub)set F of a universe X is defined thanks to a membership function denoted by μ_F which maps every element x of X into a degree $\mu_F(x)$ in the unit interval $[0, 1]$. When the degree equals 0, x does not belong at all to F, if it is 1, x is a full member of F and the closer $\mu_F(x)$ to 1 (resp. 0), the more (resp. less) x belongs to F. Clearly, a regular set is a special case of a fuzzy set where the values taken by the membership function are restricted to the pair $\{0,\ 1\}$. Beyond the intrinsic values of the degrees, the membership function offers a convenient way for ordering the elements of X and it defines a symbolic-numeric interface. The α level-cut of a fuzzy set F is defined as the (regular) set of elements whose degree of membership is greater than or equal to α and this concept bridges fuzzy sets and ordinary sets.

Similarly to a set A which is often seen as a predicate (namely, the one appearing in the intensional definition of A), a fuzzy set F is associated with a gradual (or fuzzy) predicate. For instance, if the membership function of the fuzzy set *young* is given by: $\mu_{young}(x) = 0$ for any $x \geq 30$, $\mu_{young}(x) = 1$ for any $x < 21$, $\mu_{young}(21) = 0.9$, $\mu_{young}(22) = 0.8$, ... , $\mu_{young}(29) = 0.1$, it is possible to use the predicate *young* to assess the extent to which Tom, who is 26 years old, is young ($\mu_{young}(26) = 0.4$).

The operations valid on sets (and their logical counterparts) have been extended to fuzzy sets. Their definition assumes the validity of the commensurability principle between the concerned fuzzy sets. It has been shown that it is impossible to maintain all of the properties of the Boolean algebra when fuzzy sets come into play. Fuzzy set theory starts with a strongly coupled definition of union and intersection which rely on triangular norms ($\top$) and co-norms ($\bot$) tied by de Morgan's laws. Then:

$$\mu_{A\cap B}(x) = \top(\mu_A(x),\, \mu_B(x)) \qquad \mu_{A\cup B}(x) = \bot(\mu_A(x),\, \mu_B(x))$$

The complement of a fuzzy set F, denoted by $\bar{F}$, is a fuzzy set such that: $\mu_{\bar{F}}(x) = neg(\mu_F(x))$, where *neg* is a strong negation operator and the complement to 1 is generally used. The conjunction and disjunction operators are the logical counterpart of intersection and union while the negation is the counterpart of the complement.

In practice, minimum and maximum are the most commonly used norm and co-norm because they have numerous properties among which:

- the satisfaction of all the properties of the usual intersection and union (including idempotency and double distributivity), except excluded-middle and non-contradiction laws,
- they still work with an ordinal scale, which is less demanding than numerical values over the unit interval,
- the simplicity of the underlying calculus.

[Zad65] L. Zadeh, "Fuzzy sets", *Information and Control 8*, 1965, p. 338–353.

Once these three operators given, others can be extended to fuzzy sets, such as the difference:

$$\mu_{E-F}(x) = \top(\mu_E(x),\, \mu_{\bar{F}}(x))$$

and the Cartesian product:

$$\mu_{E\times F}(x, y) = \top(\mu_E(x),\, \mu_F(y)).$$

The inclusion can be applied to fuzzy sets in a straightforward way: $E \subseteq F \Leftrightarrow \forall x, \mu_E(x) \leq \mu_F(x)$, but a gradual view of the inclusion can also be introduced. The idea is to consider that E may be more or less included in F. Different approaches can be considered, among which one is based on the notion of a fuzzy implication (the usual logical counterpart of the inclusion). The starting point is the following definition valid for sets:

$$E \subseteq F \Leftrightarrow \forall x, x \in E \Rightarrow x \in F$$

which becomes :

$$deg(E \subseteq F) = \top_x(\mu_E(x) \Rightarrow_f \mu_F(x))$$

where $\Rightarrow_f$ is a fuzzy implication whose arguments and result take their value in the unit interval. Different families of such implications have been identified (notably R-implications and S-implications) and the most common ones are:

- Kleene-Dienes implication : $a \Rightarrow_{K-D} b = \max(1 - a,\, b)$,
- Rescher-Gaines implication: $a \Rightarrow_{R-G} b = 1$ if $a \leq b$ and 0 otherwise,
- Gödel implication : $a \Rightarrow_{Go} b = 1$ is $a \leq b$ and b otherwise,
- Łukasiewicz implication : $a \Rightarrow_{Lu} b = \min(1,\, 1 - a + b)$.

Of course, fuzzy sets can also be combined in many other ways, for instance using mean operators, which do not make sense for classical sets.

3.2 Possibility theory

Possibility theory is a theory of uncertainty which aims at assessing the realization of events. The main difference with the probabilistic framework lies in the fact that it is mainly ordinal and it is not related with frequency of experiments. As in the probabilistic case, a measure (of possibility) is associated with an event. It obeys the following axioms [Zad78]:

- $\Pi(X) = 1$,
- $\Pi(\oslash) = 0$,
- $\Pi(A \cup B) = \max(\Pi(A),\, \Pi(B))$,

[Zad78] L. Zadeh, "Fuzzy sets as a basis for a theory of possibility", *Fuzzy Sets and Systems 1*, 1978, p. 3–28.

where X denotes the set of all events and A, B are two subsets of X. If $\Pi(A)$ equals 1, A is completely possible (but not certain), when it is 0, A is completely impossible and the closer to 1 $\Pi(A)$, the more possible A. From the last axiom, it appears that the possibility of $\bar{A}$, the opposite event of A, cannot be calculated from the possibility of A. The relationship between these two values (for Boolean events) is:

$$max(\Pi(A),\, \Pi(\bar{A})) = 1$$

which stems from the first and third axioms (where B is replaced by $\bar{A}$).

In other words, if A is completely possible, nothing can be deduced for $\Pi(\bar{A})$. This state of fact has led to introduce a complementary measure (N), called necessity, to assess the certainty of A. $N(A)$ is based on the fact that A is all the more certain as $\bar{A}$ is impossible [DP80]:

$$N(A) = 1 - \Pi(\bar{A})$$

and the closer to 1 $N(A)$, the more certain A. From the third axiom on possibility, one derives:

$$N(A \cap B) = \min(N(A),\, N(B)))$$

and, in general:

- $\Pi(A \cap B) \leq \min(\Pi(A),\, \Pi(B))$,
- $N(A \cup B) \geq \max(N(A),\, N(B))$.

In the possibilistic setting, a complete characterization of an event requires the computation of two measures: its possibility and its certainty. It is interesting to notice that the following property holds:

$$\Pi(A) < 1 \Rightarrow N(A) = 0.$$

It indicates that if an event is not completely possible, it is excluded that it is somewhat certain, which makes it possible to define a total order over events: first, the events which are somewhat possible but not at all certain (from ($\Pi = N = 0$ to $\Pi = 1$ and $N = 0$), then those which are completely possible and somewhat certain (from $\Pi = 1$ and $N = 0$ to $\Pi = N = 1$). This favorable situation (existence of a total order) is valid for usual events, but if fuzzy ones are taken into account, this is no longer true (because A $\cup$ $\bar{A} = X$ is not true in general when A is a fuzzy set) and the only valid property is: $\forall$ A, $\Pi(A) \geq$ N(A).

The notion of a possibility distribution [Zad78], denoted by π, plays a role similar to that of a probability distribution. It is a function from the referential X into the unit interval and:

$$\forall A \subseteq X, \Pi(A) = \sup_{x \in A} \pi(x)$$

In order to comply with the second axiom above, a possibility distribution must be such that there exists (at least) an element x_0 of X for which $\pi(x_0) = 1$. Indeed, a possibility

[DP80] D. Dubois, H. Prade, *Fuzzy set and systems: theory and applications*, Academic Press, 1980.

[Zad78] L. Zadeh, "Fuzzy sets as a basis for a theory of possibility", *Fuzzy Sets and Systems 1*, 1978, p. 3–28.

distribution can be seen as a normalized fuzzy set F which represents the knowledge about a given variable. The following formula:

$$\pi(x = a) = \mu_F(a)$$

which is often used, tells that the possibility that the actual value of the considered variable x is a, equals the degree of membership of a to the fuzzy set F. For example, Paul's age may be only imprecisely known as "close to 20", where a given fuzzy set is associated with this fuzzy linguistic expression.

3.3 Fuzzy sets, possibility theory and databases

The project is situated at the crossroads of databases and fuzzy sets. Its main objective is to broaden the capabilities offered by DBMSs according to two orthogonal lines in order to separate two distinct problems:

- flexible queries against regular databases so as to provide users with a qualitative result made of ordered elements,
- Boolean queries addressed to databases containing imprecise attribute values.

Once these two aspects solved separately, the joint issue of flexible queries against databases containing imprecise attribute values will also be considered. This can be envisaged because of the compatibility between the semantics of grades (preferences) in both fuzzy sets and possibility distributions.

It turns out that fuzzy sets offer a very convenient way for modeling gradual concepts and then flexible queries. It has been proven [BP92] that many *ad hoc* approaches (e.g., based on distances) were special cases of what is expressible using fuzzy set theory. This framework makes it possible to express sophisticated queries where the semantic choices of the user can take place (e.g., the meaning of the terms or the compensatory interaction desired between the various fuzzy conditions of a query). The works conducted in Shaman aim at extending algebraic as well as user-oriented query languages in both the relational and the object-oriented (extended relational in practice) settings. The relational algebra has already been revised in order to introduce flexible queries and a particular focus has been put on the division operation. Current works are oriented towards:

- bipolar fuzzy queries (including two parts: one viewed as a constraint, the other as a wish),
- the use of a predefined fuzzy vocabulary (which raises the question of its adequacy wrt to the actual content of the database),
- fuzzy extensions of Skyline queries (based on Pareto order),
- implementation and query optimization issues.

[BP92] P. Bosc, O. Pivert, "Some approaches for relational databases flexible querying", *Journal of Intelligent Information Systems 1*, 1992, p. 323–354.

As to possibility distributions, they are used to represent imprecise (imperfect) data. By doing so, a straightforward connection can be established between a possibilistic database and regular ones. Indeed, a possibilistic database is nothing but a weighted set of regular databases (called worlds), obtained by choosing one candidate in every distribution appearing in any tuple of every possibilistic relation. According to this view, a query addressed to a possibilistic database has a natural semantics. However, it is not realistic to process it against all the worlds due to their huge number. Then, the question tied to the querying of a possibilistic database bears mainly on the efficiency, which imposes to obviate the combinatory explosion of the worlds. The objective of the project is to identify different families of queries which comply with this requirement in the context of the relational setting, even if the initial model must obviously be extended (in particular to support imprecise data).

3.4 Query rewriting using views

3.4.1 Data integration

Information integration is the problem of combining information residing at disparate sources and providing the user with a unified view of that information. This problem has been a long standing challenge for the database community.

Two main approaches for information integration have been proposed. In the first approach, namely materialization or warehousing, data are periodically extracted from the sources and stored in a centralized repository, called a (data) warehouse. User queries are posed and executed at the warehouse with no need to access the remote information sources. Such an approach is useful in the context of intra-enterprise integration with few remote sources to integrate. It is, however, not feasible in open environments like the Web where the number of sources may be very large and dynamic.

In the second approach, called mediation or virtual integration, data stay at the sources and are collected dynamically in response to user queries [Len02,Hal03]. Mediation architectures are based on the mediator/wrapper paradigm where native information sources are *wrapped* into logical views through which the underlying sources may be accessed. The views are stored in the mediator component which additionally contains an integrated global schema that provides a single entry point to query the available information sources. The global schema acts as an interface between the user queries and the sources, freeing the users from the problem of source location and heterogeneity issues. In such an architecture, user queries posed on the global schema are rewritten in terms of logical views and then sent to the remote sources.

Briefly stated, two main approaches of mediation have been investigated [Hal01]: the GAV (Global As View) approach where the global schema is expressed as a set of views over the data sources, and the LAV (Local As View) approach where the data sources are defined as views over the global schema. Query processing is expected to be easier in the GAV approach

[Len02] M. Lenzerini, "Data Integration : A Theoretical Perspective", *in : PODS*, Madison, Wisconsin, 2002.

[Hal03] A. Halevy, "Data Integration : A status Report", *in : German Database Conference BTW-03*, Leipzig, Germany, 2003. Invited Talk.

[Hal01] A. Y. Halevy, "Answering queries using views: A survey", *VLDB Journal 10*, 4, 2001, p. 270–294.

as it can be achieved by a kind of unfolding of original queries. However, this approach suffers from a lack of extensibility as changing or adding new sources affects the global schema. On the contrary, the LAV approach is known to be highly extensible in the sense that source changes do not impact the global schema. However, in the context of the LAV approach, query processing is known to be more challenging.

A centralized mediation approach has several drawbacks including scalability, flexibility, and availability of information sources. To cope with such limitations, a new decentralized integration approach, based on a Peer-to-Peer (P2P) architecture, has been proposed. A P2P data management system [HIM+04] enables sharing heterogeneous data in a distributed and scalable way. Such a system is made of a set of peers each of which is an entire data source with its own distinct schema. Peers interested in sharing data can define pairwise mappings between their schemas. Users formulate queries over a given peer schema then a query answering system exploits relevant mappings to reformulate the original query into set of queries that enable to retrieve data from other peers.

3.4.2 Query answering in information integration systems

The problem of answering queries in mediation systems has been intensively investigated during the last decade. In particular, the investigation of this problem in the context of a LAV approach led to a great piece of fundamental theory. Recent works show that query processing in data integration is related to the general problem of answering queries using views [Hal01, Len02]. In such a setting, the semantics of queries can be formalized in terms of certain answers [AD98]. Intuitively, a certain answer to a query Q over a global (mediated) schema with respect to a set of source instances is an answer to Q in any database over the global schema that is consistent with the source instances. Therefore, the problem of answering queries in LAV-based mediation systems can be formalized as the problem of computing all the certain answers to the queries. As shown recently, this problem has a strong connection with the problem of query answering in database with incomplete information under constraints.

One of the common approaches to effectively computing query answers in mediation systems is to reduce this problem into a query rewriting problem, usually called *query rewriting using views* [Hal01,Len02,TH04]. Given a user query expressed over the global (or a peer) schema, the data sources that are relevant to answer the query are selected by means of a rewriting algorithm that allows to reformulate the user query into an equivalent or maximally subsumed (contained) query whose definition refers only to source descriptions.

The problem of rewriting queries in terms of views has been intensively investigated in the last decade (see [Hal01,Len02] for a survey). Existing research works differ w.r.t. the languages used to express a global schema, views and queries as well as w.r.t. the type of rewriting

[HIM+04] A. Y. Halevy, Z. G. Ives, J. Madhavan, P. Mork, D. Suciu, I. Tatarinov, "The Piazza Peer Data Management System.", *IEEE Trans. Knowl. Data Eng. 16*, 7, 2004, p. 787–798.

[AD98] S. Abiteboul, O. Duschka, "Complexity of Answering Queries Using Materialized Views.", *in : PODS*, p. 254–263, 1998.

[TH04] I. Tatarinov, A. Halevy, "Efficient query reformulation in peer data management systems", *in : SIGMOD '04: Proceedings of the 2004 ACM SIGMOD international conference on Management of data*, ACM Press, p. 539–550, New York, NY, USA, 2004.

considered (i.e., maximally contained or equivalent rewriting). In a nutshell, this problem has been studied for different classes of languages ranging from various sub-languages of datalog, hybrid languages combining Horn rules and description logics to semistructured data models. Recently, the problem of rewriting queries in terms of views has been investigated in the context of P2P DBMSs [HIM+04,TH04] in order to ensure scalability in terms of the number of data sources. A few recent papers also contributed to the development of data integration systems capable of taking into account imprecision or uncertainty. Most of the works along that line use probability theory in order to capture the form of uncertainty that stems from the schema definition process, or that associated with the mere existence of data, or aim at modelling the approximate nature of the semantic links between the data sources and the mediated schema.

4 Application Domains

Flexible queries have many potential application domains. Indeed, soft querying turns out to be relevant in a great variety of contexts, such as web search engines, yellow pages, classified advertisements, image or multimedia retrieval. One may guess that the richer the semantics of stored information (for instance images or video), the more difficult it is for the user to characterize his search criterion in a crisp way, i.e., using Boolean conditions. In this kind of situation, flexible queries which involve imprecise descriptions (or goals) and vague terms, may provide a convenient means for expressing information needs.

As for uncertain data management, many potential domains could take advantage of advanced systems capable of storing and querying databases where some pieces of information are imprecise/uncertain: military information systems, automated recognition of objects in images, data warehouses where information coming from more or less reliable sources must be fused and stored, etc.

In the near future, we intend to focus on two application domains:

- Open data management. One of the challenges in web data management today is to define adequate tools allowing users to extract the data that are the most likely to fulfill all or part of their information needs, then to understand and automatically correlate these data in order to elaborate relevant answers or analyses. Open data may be of various levels of quality: they may be imprecise, incomplete, inconsistent and/or their reliability/freshness may be somewhat questionable. An appropriate data model and suitable querying tools must then be defined for dealing with the imperfection that may pervade data in this context. On the other hand, it is of prime importance to provide end-users with simple and flexible means to better understand and analyze open data. The standards of W3C offer popular languages for representing both open and structured data. Another objective is to propose analytical tools suited to these languages

[HIM+04] A. Y. Halevy, Z. G. Ives, J. Madhavan, P. Mork, D. Suciu, I. Tatarinov, "The Piazza Peer Data Management System.", *IEEE Trans. Knowl. Data Eng. 16*, 7, 2004, p. 787–798.

[TH04] I. Tatarinov, A. Halevy, "Efficient query reformulation in peer data management systems", *in : SIGMOD '04: Proceedings of the 2004 ACM SIGMOD international conference on Management of data*, ACM Press, p. 539–550, New York, NY, USA, 2004.

through the construction of RDF data warehouses, whereas fuzzy-set-based data summarization approaches should constitute an important step towards making open data more intelligible to non-expert users.

- Environmental information systems. This work will be performed in collaboration with the Biological Station based in Roscoff (Finistère). The general objective is to define an information system architecture (along with an associated "toolbox") suited to the context of marine biodiversity monitoring and environmental protection. We intend to study three main aspects:
  - definition of a data warehouse model suited to this context, capable of dealing with missing values, imprecise information (a situation which often occurs due to the way data is collected and described, through sampling campaigns and human-performed labeling, in particular), uncertain data (uncertainty is unavoidable when data are obtained by means of predictive models, for instance).
  - identification of new needs in terms of query expression: new OLAP operators suitable for the model, making it possible to handle dimensions described by fuzzy concept trees, to manage fuzzy cardinalities, possibility distributions and so on.
  - knowledge discovery: we are notably interested in exploiting a concept that comes from artificial intelligence but has not been applied in the domain of data management yet: that of an analogical proportion, which underlies propositions of the type "A is to B as C is to D". We believe that discovering such "regularities" in a dataset could prove very useful for many purposes connected to environmental monitoring issues, in particular when it comes to predict the evolution of an ecosystem or the population of a species, etc.

5 Software

- PostgreSQLF is a flexible querying prototype that aims at evaluating fuzzy queries addressed to regular databases. It is an extension of PostgreSQL which implements the fuzzy query language SQLf defined in the team. This prototype is coupled with a graphical interface names ReqFlex [SPG13] that makes it easy for an end user to specify his/her fuzzy queries.

- CORTEX (CORrelaTion-based Query EXpansion): Retrieving data from large-scale databases sometimes leads to plethoric answers especially when queries are underspecified. To overcome this problem, we proposed an approach which strengthens the initial query by adding new predicates (cf. Subsection 6.2.4). These predicates are selected among predefined ones principally according to their degree of semantic correlation with the initial query. This way, we avoid an excessive modification of its initial scope. Considering the size of the initial answer set and the number of expected results specified by the

[SPG13] G. Smits, O. Pivert, T. Girault, "ReqFlex: Fuzzy Queries for Everyone", *PVLDB 6*, 12, 2013, p. 1206–1209.

user, fuzzy cardinalities are used to assess the reduction capability of these correlated predefined predicates. This approach has been implemented as a research prototype, named CORTEX, to query a database containing 10,000 ads about second hand cars [BHPS10].

- LUCIFER (Leveraging Unveiled Conflicts In FlexiblE Requests): This prototype deals with conjunctive fuzzy queries that yield an empty or poorly satisfactory answer set. It implements a cooperative answering approach which efficiently retrieves the minimal failing subqueries of the initial query (which can then be used to explain the failure and revise the query) [PSHJ12].

- FALSTAFF (FAceted search engine Leveraging Summaries of daTA with Fuzzy Features): Faced with the difficulty of formulating precise queries to retrieve items from large scale databases, interactive interfaces implementing a faceted search strategy help the users navigate through the data by successively selecting facet-value pairs. This prototype uses a faceted search strategy to construct fuzzy queries. The interactive query construction process relies on precomputed metadata that informs about the data distribution over a predefined vocabulary [SP12].

- COKE (COnnected KEywords): Keyword queries have emerged as the most convenient way to query data sources especially for unexperienced users. Introduced initially for document retrieval on the web, such queries are defined as an enumeration of keywords corresponding to a rough description of what users are looking for. The interpretation process of keyword queries has then been adapted to handle structured data like relational databases or XML documents. Instead of considering queries as an unstructured enumeration of keywords, the approach underlying the COKE system lets users structure their keyword queries using simple but meaningful grammatical connectors. Using the data structure intensively, a COKE query is translated into SQL to retrieve exact answers. An autocompletion strategy is also proposed to help users take advantage of connectors in their keyword queries [SPJP13]. An experimentation shows that the COKE system efficiently retrieves more relevant and precise answers than classical queries made of keywords enumerations and offers a good coverage of possible query patterns.

[BHPS10] P. Bosc, A. Hadjali, O. Pivert, G. Smits, "CORTEX — CORrelaTion-based query EXpansion", *in : Actes des 26e Journées Bases de Données Avancées (BDA'10), session démonstration*, 2010.

[PSHJ12] O. Pivert, G. Smits, A. Hadjali, H. Jaudoin, "LUCIFER : Un système de détection de conflits dans les requêtes flexibles", *in : Actes de la 12e Conférence Internationale Francophone sur l'Extraction et la Gestion des Connaissances (EGC'12)*, p. 617–620, 2012.

[SP12] G. Smits, O. Pivert, "A Fuzzy-Summary-Based Approach to Faceted Search in Relational Databases", *in : Proc. of the 16th East-European Conference on Advances in Databases and Information Systems (ADBIS'12)*, T. Morzy, T. Haerder, R. Wrembel (editors), *LNCS*, *7503*, Springer, p. 357–370, 2012.

[SPJP13] G. Smits, O. Pivert, H. Jaudoin, F. Paulus, "An Autocompletion Mechanism for Enriched Keyword Queries to RDF Data Sources", *in : Proc. of the 10th International Conference on Flexible Query Answering Systems (FQAS'13)*, 2013.

6 New Results

6.1 Possibilistic database modeling and querying

Participants: Olivier Pivert.

Many works have been undertaken in the area of "fuzzy databases" in the last twenty years. This term is sometimes misused or misleading since it covers both fuzzy querying against regular databases and the handling of databases that are pervaded with imprecision or uncertainty in the data (as opposed to queries).

In [22], we deal with the evaluation of aggregates queries in the framework of an uncertain database model where the notion of necessity (from possibility theory) is used to qualify the certainty that an ill-known piece of data takes a given value or belongs to a given subset. Two facets of the problem are considered, that correspond to: i) the nature of the data (certain or uncertain), and ii) the nature of the query (crisp or fuzzy) that specifies the relation over which the aggregate has to be computed.

In [24], skyline queries in the certainty-based database model are investigated. In this framework, skyline queries aim at computing the extent to which any tuple from a given relation is certainly not dominated by any other tuple from that relation.

In [23], we consider the case where information comes from multiple sources that may be conflicting. This makes uncertain the answers of a query to a set of sources. Possibility theory-based approaches to the handling of uncertainty in databases have been proposed and developed for a long time, in the case of a unique source of information. A multiple source counterpart of possibility theory has been recently proposed. Possibility and necessity set functions are then valued in terms of (possibly fuzzy) subsets of sources. Uncertainty may be assessed here either in terms of global reliability levels of sources or of tuples inside a source. When each source contains precise and certain information, each tuple is then associated with the subset of sources that supports it as being an answer to a considered query, and with the subset of sources according to which the tuple is not an answer. In fact, these subsets of sources are fuzzy as they reflect the reliability levels. When sources may contain uncertain, imprecise or missing information, there is another subset of sources that only support that it is possible that a considered tuple is an answer. The paper discusses how to rank the answers in each case. The benefit of the approach is to rank-order the answers to a query on a qualitative basis, in terms of subsets of sources and reliability levels.

6.2 Flexible querying of classical databases

6.2.1 Preference queries

Participants: Olivier Pivert, Grégory Smits, Virginie Thion, Ludovic Liétard, Daniel Rocacher.

The works presented hereafter deal with different aspects of preference queries (fuzzy and others) in a database context.

- *Survey about database preference queries.* In database research, the last two decades have witnessed a growing interest in preference queries on the one hand, and uncertain

databases on the other hand. Approaches to database preference queries may be classified into two categories according to their qualitative or quantitative nature. In the latter, preferences are expressed quantitatively by a monotone scoring function, and the overall score is positively correlated with partial scores. In the qualitative approach, preferences are defined through binary preference relations. In [3], these two families of approaches are presented and compared through some of their typical representatives.

- *Bipolar fuzzy queries and Qualitative Choice Logic.* The concept of bipolar queries is a particular way to integrate preferences inside queries where mandatory preferences, called constraints, are distinguished from optional preferences, called wishes. Constraints and wishes are respectively defined by a set of acceptable values and a set of desired values. Tuples satisfying the constraints and the wishes are returned in priority to the user. If such answers do not exist, tuples satisfying only the constraints are delivered. On the other hand, Qualitative Choice Logic (QCL) is devoted to a logic expressing preferences for Boolean alternatives. In [19], we lay down the first stones to extend QCL to fuzzy alternatives. In particular, some relationships between QCL and the bipolar expression of preferences queries are emphasized. A new type of bipolar conditions is defined in the Boolean context to express QCL statements. This new type of bipolar conditions is then made gradual and it is shown that this extension can be the basis to the definition of a fuzzy QCL model. In [20], bipolar queries are extended in order to express complex, stratified, wish conditions.

- *Skyline refinements.* Skyline queries are a popular and powerful paradigm for extracting interesting objects from a d-dimensional dataset. They rely on Pareto dominance principle to identify the skyline objects, i.e., the set of incomparable objects which are not dominated by any other object from the dataset. In [17, 16], an approach is proposed, that aims at reducing the impact of exceptional points when computing skyline queries. The phenomenon that one wants to avoid is that noisy or suspect elements "hide" some more interesting answers just because they dominate them in the sense of Pareto. The approach proposed is based on the fuzzy notion of typicality and makes it possible to distinguish between genuinely interesting points and potential anomalies in the skyline obtained. In [18], we deal with skyline queries in the context of "dirty databases", i.e., databases that may contain bad quality or suspect data. We assume that each tuple or attribute value of a given dataset is associated with a quality level and we define several extensions of skyline queries that make it possible to take data quality into account when checking whether a tuple is dominated by another. This leads to the computation of different types of gradual (fuzzy) skylines.

- *Graph databases.* In [28], we describe a fuzzy query algebra that makes it possible to handle graph databases in a flexible way. This algebra, based on fuzzy set theory and the concept of a fuzzy graph, is composed of a set of operators that can be used to express preference queries on graphs that may be fuzzy or not. The preferences that are expressible in this framework concern i) the content of the vertices of the graph and ii) the structure of the graph (which may include weighted edges when the graph is fuzzy). In [27], we build on [28] and define a query language named FUDGE, which is a fuzzy

extension of the graph database language Cypher.

- *New fuzzy comparators.* In [21], novel operators are introduced in order to compare a scalar s with a fuzzy set F. Using such comparators, one may express conditions as $s \geq_f F$, $s \leq_f F$, $s >_f F$ or $s <_f F$, that may be linguistically expressed as "s is at least F", "s is at most F", "s is more than F" and "s is less than F". In a context of database flexible querying, such operators improve the expressivity of fuzzy queries and make it possible to differentiate between values not satisfying a predicate that are "on the right" or "on the left" of this predicate.

6.2.2 Cooperative answering to flexible database queries

Participants: Grégory Smits, Olivier Pivert, Hélène Jaudoin.

The practical need for endowing information systems with the ability to exhibit cooperative behavior (thus making them more "intelligent") has been recognized at least since the early 90s. The main intent of cooperative systems is to provide correct, non-misleading and useful answers, rather than literal answers to user queries. Different aspects of this problem are tackled in the works presented hereafter.

- *Towards a cluster-based data model.* Using linguistic fuzzy variables to describe data improves the interpretability of data querying systems and thus their quality, under the condition that the considered modalities induce an indistinguishability relation in adequacy with the underlying data structure. In [31, 30], we propose a method to identify and split too general modalities so as to finally obtain a more appropriate vocabulary with respect to the data structure.

- *Empty and plethoric sets of answers.* Querying large-scale databases often leads to plethoric answers, even when fuzzy queries are used. To overcome this problem, we propose in [26] to strengthen the initial query with additional predicates, selected among predefined ones according mainly to their degree of semantic relationship with the initial query. In this approach, related predicates are identified by mining a repository of previously executed queries.

- *Association-based retrieval of similar objects.* [25] deals with the issue of extending the scope of a user query in order to retrieve objects which are similar to its "strict answers". The approach proposed exploits associations between database items, corresponding, e.g., to the presence of foreign keys in the database schema. Fuzzy concepts such as typicality, similarity and linguistic quantifiers are at the heart of the approach and make it possible to obtain a ranked list of similar answers.

6.3 Flexibility issues in data integration systems

Participants: François Goasdoué, Hélène Jaudoin, Olivier Pivert, Grégory Smits.

- *Enriched keyword queries.* In [29], we present AGGREGO SEARCH, a novel keyword-based query solution intended to help end users retrieve precise answers from semantic data sources. Contrary to existing approaches, AGGREGO SEARCH suggests grammatical connectors from natural languages during the query formulation step in order to specify the meaning of each keyword, thus leading to a complete and explicit definition of the intent of the search. An example of such a query is name of person at the head of company and author of article about "business intelligence". In order to help users formulate such connected keywords queries, a specific autocompletion strategy has been developed. A translation of the user keyword query into SPARQL is performed on-the-fly during the interactive query construction process. For this demonstration, we show how AGGREGO SEARCH has been integrated on top of a mediation system to let users intuitively define explicit and precise keyword queries in order to extract knowledge distributed in heterogeneous large semantic data sources.

- *Cloud-based RDF data management.* Cloud computing has been massively adopted recently in many applications for its elastic scaling and fault-tolerance. At the same time, given that the amount of available RDF data sources on the Web increases rapidly, there is a constant need for scalable RDF data management tools. In [4], we start by providing a compact survey of the existing works which have aimed at storing and querying large volumes of RDF data in a cloud. The second part of the chapter focuses on our specific work in this context, namely, an architecture for storing RDF data within the Amazon Web Services (AWS) cloud. Since in a cloud environment, the total consumption of storage and computing resources translates into monetary costs, reducing these costs is an objective potentially just as important as reducing response time. We present the architecture we devised to store, index and query RDF data within the AWS cloud, by exploiting the various services it provides. At the core of our proposed architecture are RDF indexing strategies which allow to guide queries directly to a (hopefully tight) superset of the RDF datasets which may provide answers to a given query, thus reducing the total work entailed by query execution. We provide a set of experiments validating the interest and performance of this architecture. This second part is based on our previous publication, however, the platform and write-up have undergone significant changes and extensions since.

- *Efficient RDF query answering techniques.* Reformulation-based query answering is a query processing technique aiming at answering queries against data, under constraints. It consists of reformulating the query based on the constraints, so that evaluating the reformulated query directly against the data (i.e., without considering any more the constraints) produces the correct answer set. In [8], we consider optimizing reformulation-based query answering in the setting of ontology-based data access, where SPARQL conjunctive queries are posed against RDF facts on which constraints expressed by an RDF Schema hold. The literature provides solutions for various fragments of RDF, aiming at computing the equivalent union of maximally-contained conjunctive queries w.r.t. the constraints. However, in general, such a union is large, thus it cannot be efficiently processed by a query engine. Our contribution is (i) to generalize the query reformulation language so as to investigate a space of reformulated queries (instead of having a sin-

gle possible choice), and then (ii) to select the reformulated query with lower estimated evaluation cost. Our experiments show that our technique enables reformulation-based query answering where the state-of-the-art approaches are simply unfeasible, while it may decrease their costs by orders of magnitude in other cases.

- *Parallel RDF data management.* RDF is an increasingly popular data model for many practical applications, leading to large volumes of RDF being created and exploited. Efficient RDF data management methods are crucial to allow applications to scale. In [14], we showcase the high efficiency of CliqueSquare, an RDF data management system designed on top of a MapReduce-like infrastructure. We demonstrate three major aspects of CliqueSquare: (*i*) significantly reducing the network traffic during query evaluation, (*ii*) evaluating even large queries into few MapReduce jobs, (*iii*) improving query performance by producing DAG-shaped plans. In all demonstration scenarios, the audience is invited to interact with the system to ask queries, explore and control the features of the platform's many optimization algorithms, and finaly select and monitor the evaluation of plans in two available clusters. A system based on these concepts, named FactMinder, devoted to fact checking on the web, is presented in [15].

- *RDF analytics.* The development of Semantic Web (RDF) brings new requirements for data analytics tools and methods, going beyond querying to semantics-rich analytics through warehouse-style tools. In [9], we fully redesign, from the bottom up, core data analytics concepts and tools in the context of RDF data, leading to the first complete formal framework for warehouse-style RDF analytics. Notably, we define (i) analytical schemas tailored to heterogeneous, semantics-rich RDF graph, (ii) analytical queries which (beyond relational cubes) allow flexible querying of the data and the schema as well as powerful aggregation and (iii) OLAP-style operations. Experiments on a fully-implemented platform demonstrate the practical interest of our approach.

- *Querying inconsistent description logic knowledge bases.* Recently several inconsistency-tolerant semantics have been introduced for querying inconsistent description logic knowledge bases. Most of these semantics rely on the notion of a repair, defined as an inclusion-maximal subset of the facts (ABox) which is consistent with the ontology (TBox). In [5, 6], we study variants of two popular inconsistency-tolerant semantics obtained by replacing classical repairs by various types of preferred repair. We analyze the complexity of query answering under the resulting semantics, focusing on the lightweight logic DL-Lite$_R$. Unsurprisingly, query answering is intractable in all cases, but we nonetheless identify one notion of preferred repair, based upon priority levels, whose data complexity is 'only' coNP-complete. This leads us to propose an approach combining incomplete tractable methods with calls to a SAT solver. An experimental evaluation of the approach shows good scalability on realistic cases.

- *Semantic search within social data.* Social content such as social network posts, tweets, news articles and more generally web page fragments is often structured. Such social content is also frequently enriched with annotations, most of which carry semantics, either by collaborative effort or from automatic tools. Searching for relevant information

in this context is both a basic feature for the users and a challenging task. In [7], we present a data model and a preliminary approach for answering queries over such structured, social and semantic-rich content, taking into account all dimensions of the data in order to return the most meaningful results.

6.4 Analogical proportions

Participants: Hélène Jaudoin, Olivier Pivert, William Correa Beltran,.

- *Prediction of missing values.* In [13, 11], a novel approach to the prediction of null values in relational databases is proposed, based on the notion of analogical proportion. We show in particular how an algorithm initially proposed in a classification context can be adapted to this purpose. Two cases are considered: that of a transactional database (where attributes are Boolean) and that where the relation considered may involve missing values of a numerical type [10]. The experimental results obtained, even though preliminary, are encouraging since the approach yields a better precision, on average, than the classical nearest neighbors technique.

- *Lazy classification.* [12] presents a novel approach for lazy classification based on the notion of analogical proportions. Our starting point is a method from the literature based on a measure of analogical dissimilarity. Based on some observations about the effectiveness of different analogical proportion situations for classification purposes, we optimize this method, reducing considerably the processing time. These results raise some questions about the reasons of the effectiveness of the analogical approach, which are discussed in the paper.

7 Other Grants and Activities

7.1 National actions

François Goasdoué gave a tutorial about "Semantic web and open linked data" at the summer school "Masses de données distribuées" in June 2014.

François Goasdoué is involved in the following projects:

- Datalyse (Investissements d'Avenir, *Big Data / Cloud computing*, 2013–2016). This project deals with Big Data management in a cloud architecture. The consortium is made of industrial partners (Eolas – Business & Decision and Les Mousqueraires), academic partners (Inria, LIFL of Univ. Lille, LIG of Univ. Grenoble, LIRMM of Univ. Montpellier), as well as the city of Grenoble as an open data provider.

- ANR JCJC Pagoda (2013–2017). PAGODA (Practical algorithms for ontology-based data access) is a basic research project whose objective is to improve the efficiency and robustness of ontology-based data access by developing scalable algorithms for query

answering in the presence of ontologies as well as pragmatic approaches to handling inconsistent data. Partners are from LIG of Univ. Grenoble, LIRMM of Univ. Montpellier, and LRI of Univ. Paris-Sud.

Ioana Manolescu, Stamatis Zampetakis, and Alexandra Roatiş, from the team OAK (Inria Saclay), visited SHAMAN in April-May 2014. I. Manolescu presented the activities of OAK, S. Zampetakis gave a talk about "CliqueSquare: Efficient RDF Query Processing in MapReduce-like Systems", and A. Roatiş presented her work about "RDF Analytics: Lenses over Semantic Graphs".

François Goasdoué, Olivier Pivert, Grégory Smits, and Virginie Thion are involved in the DGA project ODIN (Open Data INtelligence) which started in November 2014. The other partners involved are Semsoft and INRIA Saclay. The ODIN project aims to propose a data management and business intelligence solution for big data, i.e., large-scale heterogeneous and imperfect data distributed over several sources. For doing so, we intend to conceive a data processing and multidimensional analysis chain suitable for RDF data, taking into account the data quality aspect.

7.2 International actions

- Olfa Slama, Master's student from the University of Tunis (Tunisia), co-supervised by Ludovic Liétard, spent wo month in our team from April 1 to May 30.

8 Dissemination

8.1 Teaching

Project members give lectures in different faculties of engineering, in the third cycle University curriculum: "Bases de données, gradualité et imprécision" in the speciality "Interaction Intelligente avec l'Information" of the Master's degree in computer science at University of Rennes 1, and at Enssat (third year level cursus).

8.2 Scientific activities

8.2.1 Highlights of the year

- Publication by Springer of a volume edited by O. Pivert and S. Zadrożny entitled "Flexible Approaches in Data, Information and Knowledge management" [1];
- Paper [5] accepted at AAAI'14;
- Paper [9] accepted at WWW'14;
- Alexandra Roatis defended her Ph.D. thesis [2] on September 22, 2014.

8.2.2 Program committees

François Goasdoué served as a member of the following program committee:

- 30esJournées Bases de Données Avancées (BDA 2014), Grenoble-Autrans, France, 14-17 octobre 2014.
- 21st European Conference on Artificial Intelligence (ECAI 2014), Prague, Czech Republic, August 18-22, 2014.
- 11th European Semantic Web Conference (ESWC 2014), Anissaras, Crete, Greece, May 25-29, 2014.
- IEEE International Conference on Tools with Artificial Intelligence (ICTAI 14), Limassol, Cyprus, November 10-12, 2014.

H. Jaudoin served as a member of the following program committee:

- 30esJournées Bases de Données Avancées (BDA 2014), Grenoble-Autrans, France, 14-17 octobre 2014.

L. Liétard served as a member of the following program committees:

- 29th ACM Symposium on Applied Computing (SAC 2014), Gyeongju, Korea, March 24-28, 2014.

O. Pivert served as a member of the following program committees:

- 29th ACM Symposium on Applied Computing (SAC 2014), Gyeongju, Korea, March 24-28, 2014.
- 15th International Conference on Information Processing and the Management of Uncertainty in Knowledge-Based Systems (IPMU 2014), Montpellier, France, July 15-19, 2014.
- 25th International Conference on Database and Expert Systems Applications (DEXA 2014), Munich, Germany, September 1-5, 2014.
- 21st International Symposium on Methodologies for Intelligent Systems (ISMIS 2014), Roskilde, Denmark, June 25-27, 2014.
- 8th International Conference on Scalable Uncertainty Management (SUM 2014), Oxford, UK, September 15-17, 2014.
- IEEE Symposium Series on Computational Intelligence (SSCI 2014), Orlando, Florida, USA, December 9-12, 2014.
- Rencontres Francophones sur la Logique Floue et ses Applications (LFA 2014), Cargèse, France, October 22-24, 2014.

D. Rocacher served as a member of the following program committees:

- Conférence en Recherche d'Information et Applications (CORIA'14), Nancy, France, March 19-21, 2014.
- Rencontres Francophones sur la Logique Floue et ses Applications (LFA 2014), Cargèse, France, October 23-24, 2014.

8.2.3 Editorial boards

Olivier Pivert is a member of the following editorial boards:

- Journal of Intelligent Information Systems,
- Fuzzy Sets and Systems,
- International Journal of Fuzziness, Uncertainty and Knowledge-Based Systems,

8.2.4 Steering committees

O. Pivert is as a member of the steering committee of the French-speaking conference "Rencontres Francophones sur la Logique Floue et ses Applications" (LFA).

8.2.5 International advisory boards

O. Pivert is as a member of the international advisory board of the International Conference on Flexible Query-Answering Systems (FQAS).

9 Bibliography

Major publications by the team in recent years

[1] M. Bienvenu, C. Bourgaux, F. Goasdoué, "Querying Inconsistent Description Logic Knowledge Bases under Preferred Repair Semantics", *in : Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence, July 27 -31, 2014, Québec City, Québec, Canada.*, p. 996–1002, 2014.

[2] P. Bosc, L. Liétard, O. Pivert, D. Rocacher, *Gradualité et imprécision dans les bases de données*, Ellipses, 2004.

[3] D. Colazzo, F. Goasdoué, I. Manolescu, A. Roatis, "RDF analytics: lenses over semantic graphs", *in : 23rd International World Wide Web Conference, WWW '14, Seoul, Republic of Korea, April 7-11, 2014*, p. 467–478, 2014.

[4] F. Goasdoué, K. Karanasos, Y. Katsis, J. Leblay, I. Manolescu, S. Zampetakis, "Growing triples on trees: an XML-RDF hybrid model for annotated documents", *VLDB J. 22*, 5, 2013, p. 589–613.

[5] P.Bosc, O. Pivert, D.Rocacher, "About Quotient and Division of Crisp and Fuzzy Relations", *Journal of Intelligent Information Systems 29*, 2, 2007, p. 185–210.

[6] P.Bosc, O. Pivert, "About Projection-Selection-Join Queries Addressed to Possibilistic Relational Databases", *IEEE Transactions on Fuzzy Systems 13*, 1, 2005, p. 124–139.

[7] P.Bosc, O. Pivert, "About Possibilistic Queries and their Evaluation", *IEEE Transactions on Fuzzy Systems 15*, 1, 2007, p. 439–452.

[8] O. Pivert, P. Bosc, *Fuzzy Preference Queries to Relational Databases*, Imperial College Press, London, UK, 2012.

Books and Monographs

[1] O. Pivert, S. Zadrożny (editors), *Flexible Approaches in Data, Information and Knowledge Management*, *Studies in Computational Intelligence*, *497*, Springer, Heidelberg, Germany, 2014.

Doctoral dissertations and "Habilitation" theses

[2] A. Roatis, *Traitement efficace de requêtes SPARQL avec extensions OLAP pour entrepôts RDF*, PhD Thesis, University Paris-Sud – Ecole doctorale EDIPS, September 22, 2014, supervised by Ioana Manolescu, Dario Collazzo, and François Goasdoué.

Articles in referred journals and book chapters

[3] N. Bidoit, P. Bosc, L. Cholvy, O. Pivert, M. Rousset, "Bases de données et intelligence artificielle", *in : Panorama actuel de l'intelligence artificielle : ses bases méthodologiques, ses développements*, P. Marquis, O. Papini, and H. Prade (editors), Cépaduès Editions, Toulouse, 2014.

[4] F. Bugiotti, J. Camacho-Rodríguez, F. Goasdoué, Z. Kaoudi, I. Manolescu, S. Zampetakis, "SPARQL Query Processing in the Cloud", *in : Linked Data Management*, A. Harth, K. Hose, and R. Schenkel (editors), *Emerging Directions in Database Systems and Applications*, Chapman and Hall/CRC, April 2014, `http://hal.inria.fr/hal-00909121`.

Publications in Conferences and Workshops

[5] M. Bienvenu, C. Bourgaux, F. Goasdoué, "Querying Inconsistent Description Logic Knowledge Bases under Preferred Repair Semantics", *in : Proc. of the 28th Conference of the Association for the Advancement of Artificial Intelligence (AAAI'14)*, Québec City, Canada, 2014.

[6] M. Bienvenu, C. Bourgaux, F. Goasdoué, "Querying Inconsistent Description Logic Knowledge Bases under Preferred Repair Semantics (Extended Abstract)", *in : Proc. of the 27th International Workshop on Description Logics (DL'14)*, Vienna, Austria, 2014.

[7] R. Bonaque, B. Cautis, F. Goasdoué, I. Manolescu, "Toward Social, Structured and Semantic Search", *in : Proc. of Surfacing the Deep and the Social Web (SDSW), workshop co-located with the 13th International Semantic Web Conference (ISWC 2014)*, Riva del Garda - Trentino, Italy, 2014.

[8] D. Bursztyn, F. Goasdoué, I. Manolescu, A. Roatiş, "Optimizing Reformulation-based Query Answering in RDF", *in : Actes des 30esjournées Bases de Données Avancées (BDA'14)*, Grenoble-Autrans, France, 2014.

[9] D. Colazzo, F. Goasdoué, I. Manolescu, A. Roatiş, "RDF Analytics: Lenses over Semantic Graphs", *in : Proc. of the 23rd International World Wide Web Conference (WWW'14)*, Seoul, Korea, 2014.

[10] W. Correa Beltran, H. Jaudoin, O. Pivert, "Analogical Prediction of Null Values: The Numerical Attribute Case", *in : Proc. of the 18th East-European Conference on Advances in Databases and Information Systems (ADBIS'14), LNCS vol. 8716*, p. 323–336, Ohrid, Republic of Macedonia, 2014.

[11] W. Correa Beltran, H. Jaudoin, O. Pivert, "Estimating Null Values in Relational Databases Using Analogical Proportions", *in : Proc. of the 15th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU'14), CCIS vol. 444*, p. 110–119, Montpellier, France, 2014.

[12] W. Correa Beltran, H. Jaudoin, O. Pivert, "Lazy Analogical Classification: Optimization and Precision Issues", *in : Proc. of the 8th International Conference on Scalable Uncertainty Management (SUM'14), LNAI vol. 8720*, p. 80–85, Oxford, UK, 2014.

[13] W. Correa Beltran, H. Jaudoin, O. Pivert, "Prédiction de valeurs manquantes dans les bases de données – Une première approche fondée sur la notion de proportion analogique", *in : Actes de la 14e Conférence Internationale Francophone sur l'Extraction et la Gestion des Connaissances (EGC'14)*, Rennes, France, 2014.

[14] B. Djahandideh, F. Goasdoué, Z. Kaoudi, I. Manolescu, J. Quiané-Ruiz, S. Zampetakis, "How to deal with Cliques at Work", *in : Actes des 30esjournées Bases de Données Avancées (BDA'14), Session démonstration*, Grenoble-Autrans, France, 2014.

[15] F. Goasdoué, K. Karanasos, Y. Katsis, J. Leblay, I. Manolescu, S. Zampetakis, "Fact Checking and Analyzing the Web with FactMinder", *in : Proc. of Computation + Journalism Symposium 2014*, New York, NY, USA, 2014.

[16] H. Jaudoin, O. Pivert, D. Rocacher, "Exception-Tolerant Skyline Queries", *in : Proc. of the 15th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU'14), CCIS vol. 444*, p. 120–129, Montpellier, France, 2014.

[17] H. Jaudoin, O. Pivert, D. Rocacher, "Requêtes skyline en présence d'exceptions", *in : Actes de la 14e Conférence Internationale Francophone sur l'Extraction et la Gestion des Connaissances (EGC'14)*, Rennes, France, 2014.

[18] H. Jaudoin, O. Pivert, G. Smits, V. Thion, "Data-Quality-Aware Skyline Queries", *in : Proc. of the 21st International Symposium on Methodologies for Intelligent Systems (ISMIS'14), LNAI vol. 8502*, p. 530–535, Roskilde, Denmark, 2014.

[19] L. Liétard, A. Hadjali, D. Rocacher, "Towards a Gradual QCL Model for Database Querying", *in : Proc. of the 15th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU'14)*, Montpellier, France, 2014.

[20] L. Liétard, D. Rocacher, A. Hadjali, "Les conditions multipolaires floues", *in : Actes des Rencontres Francophones sur la Logique Floue et ses Applications (LFA'14)*, Cargèse, France, 2014.

[21] P. Nerzic, G. Smits, O. Pivert, "Un nouvel usage des prédicats flous pour l'interrogation flexible de base de données", *in : Actes des Rencontres Francophones sur la Logique Floue et ses Applications (LFA'14)*, Cargèse, France, 2014.

[22] O. Pivert, H. Prade, "Dealing with Aggregate Queries in an Uncertain Database Model Based on Possibilistic Certainty", *in : Proc. of the 15th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU'14), CCIS vol. 444*, p. 150–159, Montpellier, France, 2014.

[23] O. Pivert, H. Prade, "Querying Uncertain Multiple Sources", *in : Proc. of the 8th International Conference on Scalable Uncertainty Management (SUM'14), LNAI vol. 8720*, p. 286–291, Oxford, UK, 2014.

[24] O. Pivert, H. Prade, "Skyline Queries in an Uncertain Database Model Based on Possibilistic Certainty", *in : Proc. of the 8th International Conference on Scalable Uncertainty Management (SUM'14), LNAI vol. 8720*, p. 280–285, Oxford, UK, 2014.

[25] O. Pivert, G. Smits, A. Moreau, H. Jaudoin, "Réponses connexes fondées sur des associations typiques", *in : Actes des Rencontres Francophones sur la Logique Floue et ses Applications (LFA'14)*, Cargèse, France, 2014.

[26] O. Pivert, G. Smits, "Plethoric Answers to Fuzzy Queries: A Reduction Method Based on Query Mining", *in : Proc. of the 21st International Symposium on Methodologies for Intelligent Systems (ISMIS'14), LNAI vol. 8502*, p. 295–304, Roskilde, Denmark, 2014.

[27] O. Pivert, V. Thion, H. Jaudoin, G. Smits, "On a Fuzzy Algebra for Querying Graph Databases", *in : Proc. of the IEEE International Conference on Tools with Artificial Intelligence (ICTAI'14)*, Limassol, Cyprus, 2014.

[28] O. Pivert, V. Thion, H. Jaudoin, G. Smits, "Une algèbre floue pour l'interrogation flexible de bases de données graphes", *in : Actes des 30esjournées Bases de Données Avancées (BDA'14)*, Grenoble-Autrans, France, 2014.

[29] G. Smits, O. Pivert, H. Jaudoin, F. Paulus, "AGGREGO SEARCH: Interactive Keyword Query Construction", *in : EDBT*, S. Amer-Yahia, V. Christophides, A. Kementsietsidis, M. N. Garofalakis, S. Idreos, V. Leroy (editors), OpenProceedings.org, p. 636–639, 2014.

[30] G. Smits, O. Pivert, M.-J. Lesot, "Ajustement automatique de vocabulaire expert par scission de modalité", *in : Actes des Rencontres Francophones sur la Logique Floue et ses Applications (LFA'14)*, Cargèse, France, 2014.

[31] G. Smits, O. Pivert, M.-J. Lesot, "A Vocabulary Revision Method Based on Modality Splitting", *in : Proc. of the 15th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU'14), CCIS vol. 444*, p. 140–149, Montpellier, France, 2014.